New York City Motor Vehicle Collisions

Rui Shi

I. Motivation for Dataset Creation

A. Why was the dataset created?

This data is collected because the NYC Council passed Local Law 11 in 2011. This data is manually run every month and reviewed by the TrafficStat Unit before being posted on the NYPD website.

B. What (other) tasks could the dataset be used for?

This data can be used by the public to see how dangerous/safe intersections are in NYC. The information is presented in pdf and excel format to allow the casual user to just view the information in the easy to read pdf format or use the excel files to do a more in-depth analysis.

C. Has the dataset been used for any tasks already?

Yes. This dataset has been used to analysis the influence of time factors, boroughs, and locations. Also, the dataset has been analysing along with weather data to identify the influence of weather condition.

D. Who funded the creation of the dataset?

This dataset is founded by NYC council.

II. Dataset Composition

A. What are the instances?

Each instance represents a collision in NYC by city, borough, precinct and cross street.

B. Are relationships between instances made explicit in the data (e.g., social network links, user/movie ratings, etc.)?

They are all collisions happened in NYC with close accident time.

C. How many instances of each type are there?

There are 1.46 million instances which are collisions happened in NYC from 2014 to 2019.

D. What data does each instance consist of?

Each instance consists of labels indicating the collision time, location, number of person injured or killed ,and vehicle types.

E. Is everything included or does the data rely on external resources? (e.g., websites, tweets, datasets) If external resources, a) are there guar- antees that they will exist, and remain constant, over time; b) is there an official archival version. Are there licenses, fees or rights associated with any of the data?

Everything is included.

F. Are there recommended data splits or evaluation measures? (e.g., training, development, testing; accuracy/AUC)

Unknown.

G. What experiments were initially run on this dataset? Have a summary of those results and, if available, provide the link to a paper with more information here.

The dataset was originally released without reported experimental results. But after the data released, there were some experiments did for analysing the trend of collision accidents in NYC.

III. Data Collection Process

A. How was the data collected? (e.g., hardware apparatus/sensor, manual human curation, software program, software interface/API; how were these constructs/measures/methods validated?)

This data is collected by combining data from police and from the accident and emergency department.

B. Who was involved in the data collection process? (e.g., students, crowd workers) How were they compensated? (e.g., how much were crowd workers paid?)

Unknown.

C. Over what time-frame was the data collected? Does the collection time- frame match the creation time-frame?

The data is updated every weekday.

D. How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part of speech tags; model-based guesses for age or language)? If the latter two, were they validated/verified and if so how?

The instances are depended on collisions and related accident information.

E. Does the dataset contain all possible instances? Or is it, for instance, a sample (not necessarily random) from a larger set of instances?

This dataset contains all accidents happened in NYC from 2014 to 2019, and updates every weekday.

F. If the dataset is a sample, then what is the population? What was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? Is the sample representative of the larger set (e.g., geo- graphic coverage)? If not, why not (e.g., to cover a more diverse range of instances)? How does this affect possible uses?

This dataset is not a sample.

G. Is there information missing from the dataset and why? (this does not include intentionally dropped instances; it might include, e.g., redacted text, withheld documents) Is this data missing because it was unavailable?

There are lots of location information missing. Some instances have part of the location information for example the latitude and longitude but miss the zip code.

H. Is there information missing from the dataset and why? (this does not include intentionally dropped instances; it might include, e.g., redacted text, withheld documents) Is this data missing because it was unavailable?

No data is missing.

I. Are there any known errors, sources of noise, or redundancies in the data?

IV. Data Processing

A. What preprocessing/cleaning was done? (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values, etc.)

B. Was the raw data saved in addition to the preprocessed/cleaned data? (e.g., to support unanticipated future uses)

Yes.

C. Is the preprocessing software available?

No.

D. Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet?

V. Data Distribution

A. How is the dataset distributed? (e.g., website, API, etc.; does the data have a DOI; is it archived redundantly?)

The dataset can be downloaded from new-york-city-motor-vehicle-collisions

B. When will the dataset be released/first distributed? (Is there a canonical paper/reference for this dataset?)

The dataset was first released in 2014.

C. What license (if any) is it distributed under? Are there any copyrights on the data?

Unknown.

D. Are there any fees or access/export restrictions?

No.

VI. Data maintenance

A. Who is supporting/hosting/maintaining the dataset? How does one contact the owner/curator/manager of the dataset (e.g. email address, or other contact info)?

TrafficStat Unit is maintaining the dataset.

B. Will the dataset be updated? How often and by whom? How will up- dates/revisions be documented (e.g., mailing list, GitHub)? Is there an erratum?

This dataset is updated every weekday by TrafficStat Unit. TrafficStat Unit is maintaining the dataset.

C. If the dataset becomes obsolete how will this be communicated?

This will be posted on the dataset webpage.

D. Is there a repository to link to any/all papers/systems that use this dataset?

Yes. The link is NYC Open Data

E. If others want to extend/augment/build on this dataset, is there a mechanism for them to do so? If so, is there a process for tracking/assessing the quality of those contributions. What is the process for communicating/distributing these contributions to users?

Unknown.

VII. Legal & Ethical Considerations

A. If the dataset relates to people (e.g., their attributes) or was generated by people, were they informed about the data collection? (e.g., datasets that collect writing, photos, interactions, transactions, etc.)

People who is involved in any accident where there is damage to the property of one individual that is more than $\$1,000$ are required by the NY State Vehicle and Traffic Law to file an accident report. In 2012, the New York City approved Local Law 11, requiring all public data need to be made freely available on a single web portal.

B. If it relates to other ethically protected subjects, have appropriate obligations been met? (e.g., medical data might include information collected from animals)

Not applicable.

C. If it relates to people, were there any ethical review applications/reviews/approvals? (e.g. Institutional Review Board applications)

Personal information is not included in the data. If disclosed would result in an unwarranted invasion of personal policy, the records are deniable.

D. If it relates to people, were they told what the dataset would be used for and did they consent? What community norms exist for data collected from human communications? If consent was obtained, how? Were the people provided with any mechanism to revoke their consent in the future or for certain uses?

Personal information is not included in the data. If disclosed would result in an unwarranted invasion of personal policy, the records are deniable.

E. If it relates to people, could this dataset expose people to harm or legal action? (e.g., financial social or otherwise) What was done to mitigate or reduce the potential for harm?

No. Personal information is not included in the data. If disclosed would result in an unwarranted invasion of personal policy, the records are deniable.

F. If it relates to people, were they provided with privacy guarantees? If so, what guarantees and how are these ensured?

This dataset only show collision information related location and time. No information related to people is showed. Personal information is not included in the data. If disclosed would result in an unwarranted invasion of personal policy, the records are deniable.

G. Does the dataset comply with the EU General Data Protection Regulation (GDPR)? Does it comply with any other standards, such as the US Equal Employment Opportunity Act?

This data complies with mandares wstablished in the Open Data Law.

H. Does the dataset contain information that might be considered sensitive or confidential? (e.g., personally identifying information)

No.

I. Does the dataset contain information that might be considered inappropriate or offensive?

No.